Grid Computing at TACC

Jay Boisseau, Director
Texas Advanced Computing Center
The University of Texas at Austin
August 9, 2004

Motivations

- Inform conference attendees about TACC grid computing activities
- Initiate discussions with DoD researchers to share information, accelerate efforts
- Develop collaborations / new PET activities with DoD community in grid computing

Outline

1. TACC Overview (brief)
2. TACC Grid Computing Activities
3. UT Grid Project
4. TACC Support of DoD CWO Community

TACC Overview

TACC Mission & Strategy

To enhance the research & education programs of The University of Texas at Austin and its partners through research, development, operation & support of advanced computing technologies.

TACC Mission & Strategy

To accomplish this mission, TACC:

- Evaluates, acquires & operates advanced computing systems
- Provides training, consulting, and documentation to users

Resources & Services

- Collaborates with researchers to apply advanced computing techniqu
- Conducts research & development t produce new computational technologies

Research & Developme nt

TACC Advanced Computing Technology Areas

- High Performance Computing (HPC)

  numerically intensive computing: produces data

TACC Advanced Computing Technology Areas

- High Performance Computing (HPC)
  - numerically intensive computing: produces data
- Scientific Visualization (SciVis)
  - rendering data into information & knowledge

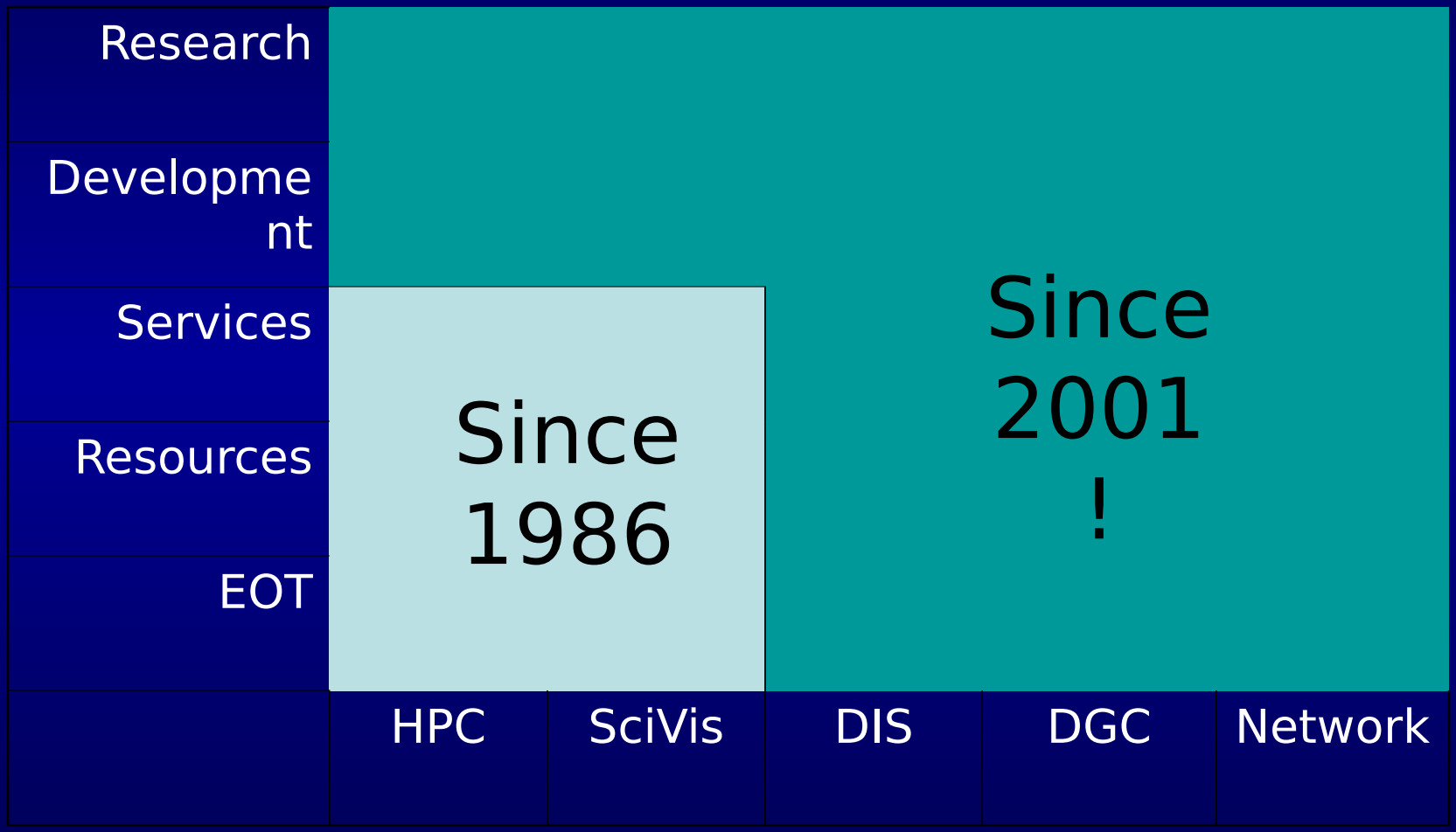
TACC Advanced Computing Technology Areas

- High Performance Computing (HPC)
  - numerically intensive computing: produces data
- Scientific Visualization (SciVis)
  - rendering data into information & knowledge
- Data & Information Systems (DIS)
  - managing and analyzing data for information & knowledge

TACC Advanced Computing Technology Areas

- High Performance Computing (HPC)
  - numerically intensive computing: produces data
- Scientific Visualization (SciVis)
  - rendering data into information & knowledge
- Data & Information Systems (DIS)
  - managing and analyzing data for information & knowledge
- Distributed and Grid Computing (DGC)
  - Integrating diverse resources, data, and people to produce and share knowledge

TACC Activities & Scope



TACC Applications Focus Areas

- TACC advanced computing technology R&D must be driven by applications
- TACC Applications Focus Areas
  - Chemistry -> Biosciences
  - Climate/Weather/Ocean -> Geosciences
  - CFD

TACC HPC & Storage Systems

LONESTAR

Cray-Dell Xeon Linux Cluster
1028 CPUs (6.3 Tflops)
1 TB memory, 40+ TB disk

LONGHORN

IBM Power4 System
224 CPUs (1.16 Tflops)
½ TB memory, 7.1 TB disk

TEJAS

IBM Linux Pentium III Cluster
64 CPUs (64 Gflops)
32 GB memory, ~1 TB disk

ARCHIVE

STK PowderHorns (2)
2.8 PB max capacity
managed by Cray DMF

SAN

Sun SANs (2)
8 TB / 4 TB
to be expanded

TACC

THE UNIVERSITY OF TEXAS AT AUSTIN

ACES VisLab

- Front and Rear Projection Systems
  - 3x1 cylindrical immersive environment, 24’ diameter
  - 5x2 large-screen, 16:9 panel tiled display
  - Full immersive capabilities with head/motion tracking

- High end rendering systems
  - Sun E25K: 128 processors, ½ TB memory, > 3 Gpoly/sec
  - SGI Onyx2: 24 CPUs, 6 IR2 Graphics Pipes, 25 GB Memory
- Matrix switch between systems, projectors, rooms

TACC Services

- TACC resources and services include:
  - Consulting
  - Training
  - Technical documentation
  - Data storage/archival
  - System selection/configuration consulting
  - System hosting

TACC R&D – High Performance Computing

- Scalability, performance optimization, and performance modeling for HPC applications
- Evaluation of cluster technologies for HPC
- Portability and performance issues of applications on clusters
- Climate, weather, ocean modeling collaboration and support of DoD
- Starting CFD activities

TACC R&D – Scientific Visualization

- Feature detection / terascale data analysis
- Evaluation of performance characteristics and capabilities of high-end visualization technologies
- Hardware accelerated visualization and computation on GPUs
- Remote interactive visualization / grid-enabled interactive visualization

TACC R&D – Data & Information Systems

- Newest technology group at TACC
- Initial R&D focused on creating/hosting scientific data collections
- Interests / plans
  - Geospatial and biological database extensions
  - Efficient ways to collect/create metadata
  - DB clusters / parallel DB I/O for scientific data

TACC R&D – Distributed & Grid Computing

- Web-based grid portals
- Grid resource data collection and information services
- Grid scheduling and workflow
- Grid-enabled visualization
- Grid-enabled data collection hosting
- Overall grid deployment and integration

TACC R&D - Networking

- Very new activities:
  - Exploring high-bandwidth (OC-12, GigE, OC-48, OC192) remote and collaborative grid-enabled visualization
  - Exploring network performance for moving terascale data on 10 Gbps networks (TeraGrid)
  - Exploring GigE aggregation to fill 10 Gbps networks (parallel file I/O, parallel database I/O)
- Recruiting a leader for TACC networking R&D activities

TACC Growth

- TACC has grown in staff, external funding, and compute capabilities by 4-50x or more!
- New infrastructure provides UT with comprehensive, balanced, world-class resources:
  - 50x HPC capability
  - 20x archival capability
  - 10x network capability
  - New SAN
  - Much enhanced (more complete) vislab
  - 8x external funding
  - 4x staff

Summary

- TACC provides terascale HPC, SciVis, storage, data collection, and network resources
- TACC provides expert support services: consulting, documentation, and training in HPC, SciVis, and Grid
- TACC conducts applied research & development in these advanced computing technologies
- TACC has become one of the leading academic advanced computing centers in years

TACC Grid Computing Activities

Everybody Uses Distributed Computing

- Consumers and IT workers use PCs to access remote web servers, financial systems, etc.
- Researchers already use distributed systems:
  - Local workstations for some development, small simulations
  - HPC in at big centers
  - Visualization back in their lab or in a VisLab
  - Archival storage to SANs, NASes, tape silos, etc.
  - Collaborate with peers at other institutions
- Grid computing enables & simplifies resource sharing, coordination

Grid Computing: My View

- Grid computing is a standard, ‘complete’ set of distributed computing software capabilities
- Grid computing must provide some basic functions
  - resource discovery and information collection & publishing
  - data management on and between resources
  - process management on an between resources
  - common security mechanism underlying the above
- No grid computing package provides everything
- ‘Open Grid Services Architecture’ (OGSA) makes it possible to build the components and make them work together

Grid Computing: My View

- There is much hype around Grid computing, but also much real-world importance
  - I prefer the term ‘grid computing’ over ‘The Grid’
- TACC focuses on Grid computing as means of
  - enhancing our HPC, SciVis, and massive data storage
  - integrating researchers’ local computing systems with ours
  - eventually, integrating research instruments for research that also requires HPC, SciVis, massive data storage

TACC Grid Tools and Deployments

- Grid Computing Tools
  - GridPort and Grid-enabled Portals
  - GridShell
  - CSF
  - GridFlow / GridSteer
- Cyberinfrastructure Deployments
  - TeraGrid
  - Texas Internet Grid for Research & Education (TIGRE)
  - UT Grid

GridPort and Grid-Enabled Portals

- Web portals simplify:
  - Information presentation from distributed sources
  - Interaction with distributed, remote resources
- Grid-enabled portals enable grid services via simple interfaces

GridShell

- *Shell:* A program that works with the operating system as a command processor, used to enter commands and initiate their execution.
- GridShell is an extension of TCSH and BASH that includes transparent distributed execution and data transfer features for intra and inter cluster execution of programs.

Distributed Shell in a Cluster

- Transparently maps and executes commands using the underlying batch queuing environment (execution transparency)
- Scripts can be run, with minimal change, across clusters running different batch queuing environments, e.g. LSF, PBS, SGE, Condor, etc. (heterogeneous)
- Overloads redirection operators to provide data transfer capabilities across nodes using gsiftp or scp. (data transparency)
- Provides the capability of spawning parallel scripting tasks using nodes allocated to it from the batch queuing environment (distributed computation)
- Provides simple message-passing capability for the distributed shell scripts to synchronize and send data across nodes in a cluster job (distributed communication)
- Provides a scripting environment to orchestrate the execution of distributed tasks in the cluster environment (distributed workflows)

GridShell Broad Design Goals

- *Execution Transparency*
- *Data Transparency*
- *Support Heterogeneity*
- *Extensible Architecture*
- *Support Local Autonomy*

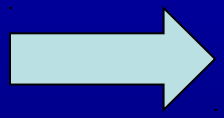
Parametric Sweep with local job queue throttling

```
#!/bin/sh
n=0
while [ $n -lt 1000 ]; do

   cat >scr$n <<EOF
   TASKID=$n
   export TASKID
   ./a.out >out${n}.dat
   EOF
   qsub scr$n
   n=`expr $n + 1`

   flag=0
   while [ $flag –eq 0 ] ;
do
      cnt=`qstat | wc –l`
      if [ $cnt –lt 27 ] ;
then
         flag=1
      else
         sleep 1
      endif
   done
done
```

OR



```
#!/bin/gtcsh

a.out in 1000 instances
```

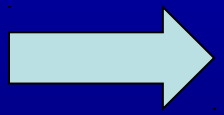
Parametric sweep across the grid

```
#!/bin/sh

cluster=`resource_broker`
localhost=`hostname`

if [ “$cluster” = “$localhost” ] then
    a.out in 1000 instances
else
    globus-url-copy a.out gsiftp://$cluster/$pwd
    globus-job-run $cluster $pwd/a.out
fi
```

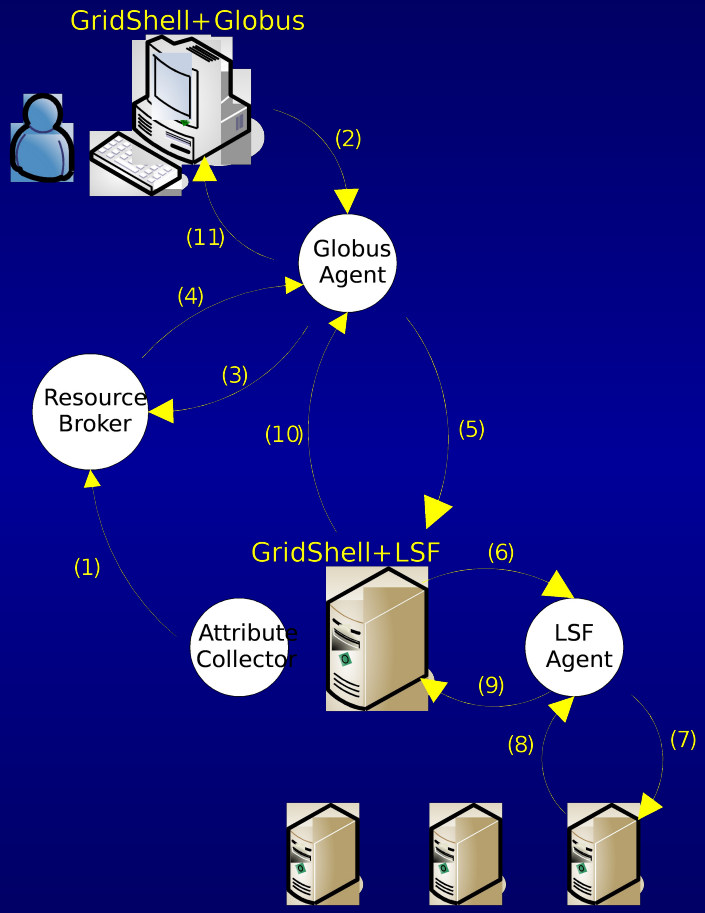
OR



```
#!/bin/gtcsh

a.out in 1000 instan
```

GridShell Grid Components



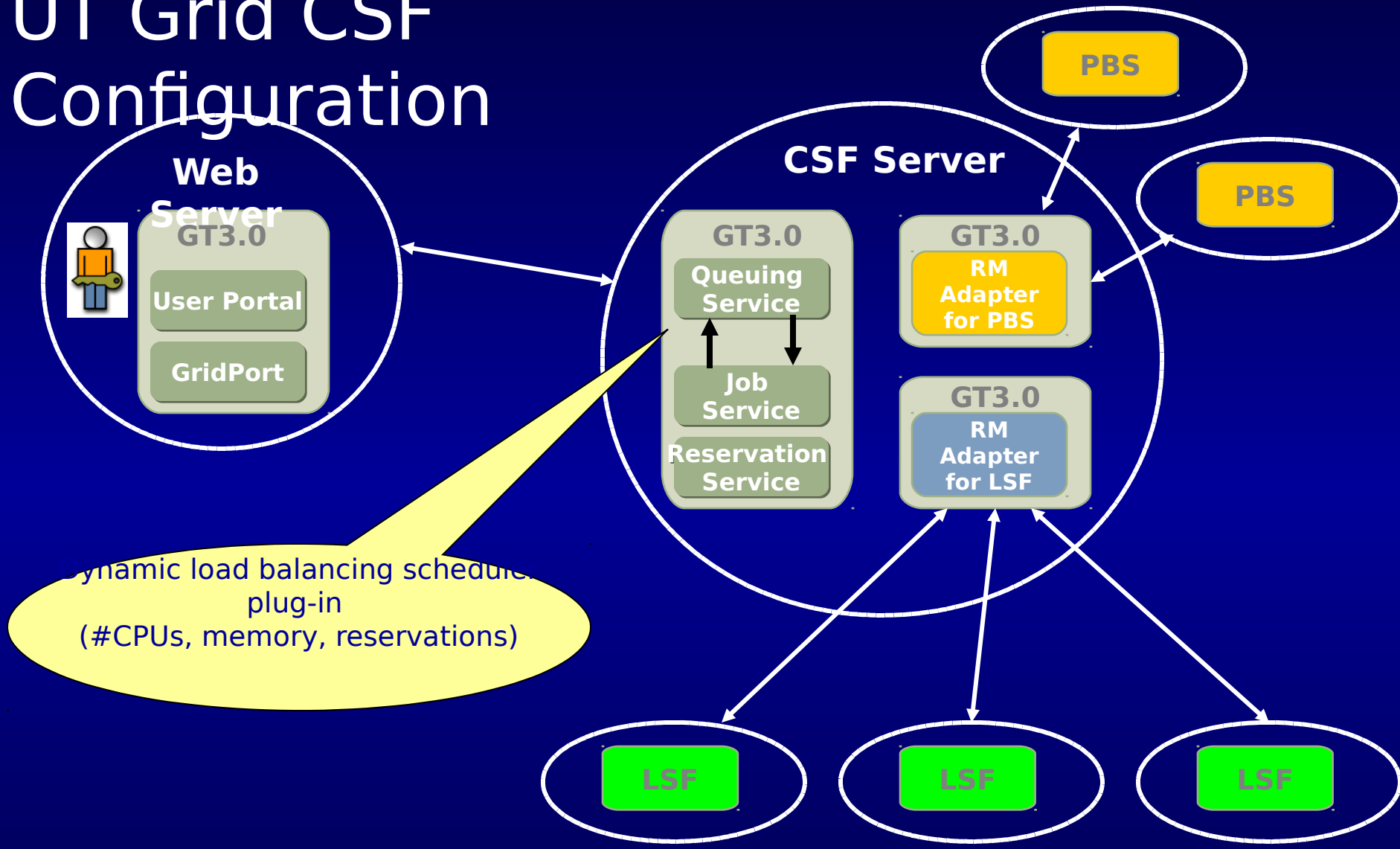
Community Scheduling Framework (CSF)

- CSF is an open-source metascheduler that allows the scheduling of jobs to multiple, heterogeneous cluster resources
  - Enables global access and co-ordination while maintaining local control and ownership of resources
  - Provides an extensible framework to enable customization of policies using scheduler “plug-ins”
  - Functionality for scheduling and management of jobs and advanced reservation and monitoring of resources
- Web Services API based on current Grid Services standards
  - Planned for inclusion in next Globus Toolkit release (version 4.x)
- Contributed by Platform Computing to open-source community
  - Distributed under the Globus Public License
  - Sourceforge link: http://sourceforge.net/projects/gcsf/

CSF Services

- Job Service
  - Creates, monitors and controls compute jobs
- Reservation Service
  - Reserves compute resources at a specified time on behalf of a job
  - Advanced reservations possible only if supported by local resource manager (e.g. LSF, MAUI)
- Queuing Service
  - Service where administrators can customize and define scheduling policies for the Virtual Organization
- Resource Manager Adapter Service
  - Provides a Grid service interface which bridges the Grid service protocol and resource managers (e.g. LSF, PBS, SGE, …)
  - RM Adapters currently available include LSF & PBS/MAUI
  - Plans to develop adapters for SGE/MAUI & Condor-G
- Global Information Service
  - Propagation of information between resource managers and the metascheduler
  - Information used for making dynamic scheduling decisions

UT Grid CSF Configuration



GridFlow & GridSteer

- Simple Globus-based tools to enable automation of compute/visualization cycle in HPC applications
  - GridFlow pipelines results from one computational process to a visualization process on a remote system
  - GridSteer directs a computational process by analyzing results from a remote process
- Functions based on Globus Toolkit
  - Globus software required on clusters
- Automation is achieved by adding GridFlow and GridSteer calls to the HPC compute code.
  - Simple application code changes are required

GridFlow: Typical Use Case

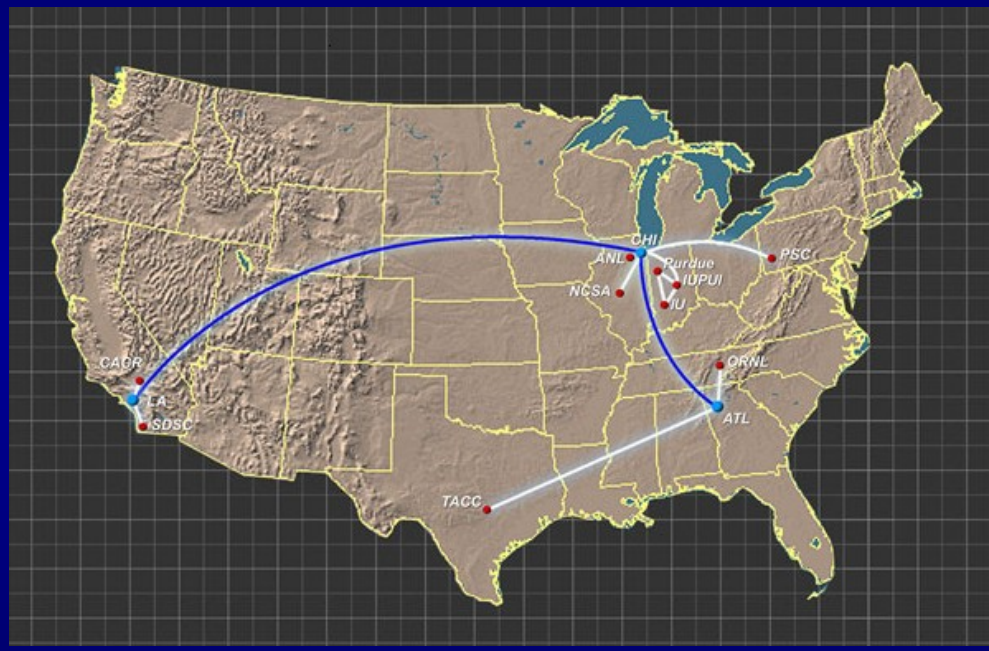
- HPC user wants to run simulations on HPC system and view results on visualization system. Approach used today:
  1. Run job on HPC system and monitor for completion
  2. Move output data to visualization system
  3. Launch post-processing visualization software
  4. View visualization images and repeat Steps 1-4, if necessary
- Using Grid Technology, this use case can be enhanced:
  1. submit HPC job to Grid scheduler to select "best" system
  2. send intermediate results to specified visualization system to render
  3. view intermediate results and determine if parameters need to be adjusted
  4. If so, update the parameters on the visualization system and have the HPC simulation automatically pick the new parameters.
- Benefits for user
  - Automation simplifies the end-to-end process
  - Quicker turnaround by detecting errors earlier in the process

GridFlow/GridSteer sample

```
read input data
read input parameters
initialize variables
do I = 1, max_timesteps
    solve equations
    write intermediate data to files every N timesteps
    call GridFlow to push intermediate data to
  visualization system
                    and generate image
     call GridSteer to get remote parameters
                    and take appropriate action based on
  return code
end do
write output data
end program
```

NSF TeraGrid: National Cyberinfrastructure for Computational Science

- TeraGrid is world's largest cyerinfrastructure for computational research
- Includes NCSA, SDSC, PSC, Caltech, Argonne, Oak Ridge, Indiana, Purdue
- Massive bandwidth! Each connection is one or more 10 Gbps links!



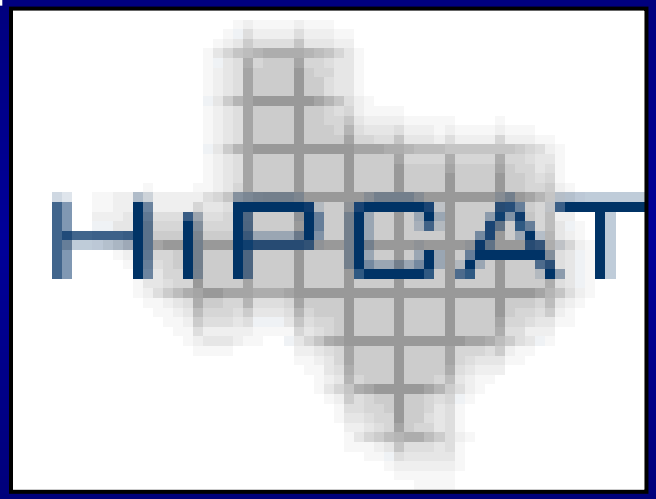
- TACC will provide terascale computing, storage,
  and visualization resources
- UT will provide terascale geosciences data sets

TACC

THE UNIVERSITY OF TEXAS AT AUSTIN

Texas Internet Grid for Research & Education (TIGRE)

- Multi-university grid: Texas, A&M, Houston, Rice, Texas Tech
  - Build-out in 2004-5
  - Will integrate additional universities
- Will facilitate academic research capabilities across Texas using Internet2 initially
- Will extend to industrial partners to foster academic/industrial collaboration on R&D



UT Grid: Enable Campus-wide Terascale Distributed Computing

- Vision: provide high-end systems, but move from ‘island’ to hub of campus computing continuum
  - provide models for local resources (clusters, vislabs, etc.), training, and documentation
  - develop procedures for connecting local systems to campus grid
    - single sign-on, data space, compute space
    - leverage every PC, cluster, NAS, etc. on campus!
  - integrate digital assets into campus grid
  - integrate UT instruments & sensors into campus grid
- Joint project with IBM

UT Grid Project

UT Grid: Initial Focus on Computing

- High-throughput parallel computing
  - Project Rodeo
  - Use CSF to schedule to LSF, PBS, SGE clusters across campus
  - Use Globus 3.2 -> GT4
- High-throughput serial computing
  - Project Roundup uses United Devices software on campus PCs
  - Also interfacing to Condor flock in CS department

UT Grid: Initial Focus on Computing

- Develop CSF adapters for popular resource management systems through collaboration:
  - LSF: done by Platform Computing
  - Globus: done by Platform Computing
  - PBS: partially done
  - SGE
  - LoadLeveler
  - Condor

UT Grid: Initial Focus on Computing

- Develop CSF capability for flexible job requirements:
  - Serial vs parallel: no diff, just specify Ncpus
  - Number: facilitate ensembles
  - Batch: whenever, or by priority
  - Advanced reservation: needed for coupling, interactive
  - On-demand: needed for urgency
- Integrate data management for jobs into CSF
  - SAN makes it easy
  - GridFTP is somewhat simple, if crude
  - Avaki Data Grid is a possibility

UT Grid: Initial Focus on Computing

- Completion time in a compute grid is a function of
  - data transfer times
    - Use NWS for network bandwidth predictions, file transfer time predictions (Rich Wolski, UCSB)
  - queue wait times
    - Use new software from Wolski for prediction of start of execution in batch systems
  - application performance times
    - Use Prophesy (Valerie Taylor) for applications performance prediction
- Develop CSF scheduling module that is data, network, and performance aware

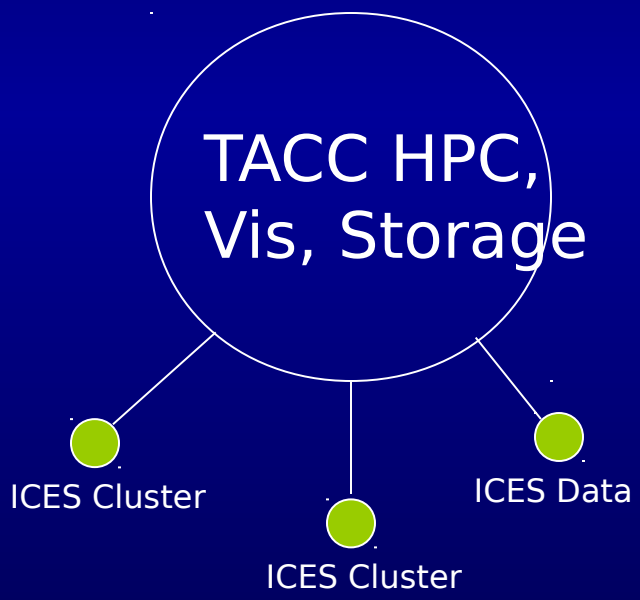
UT Grid: Logical View

- *Integrate* distributed TACC resources first (Globus, LSF, NWS, SRB, United Devices, GridPort)

(actually spread across two campuses)

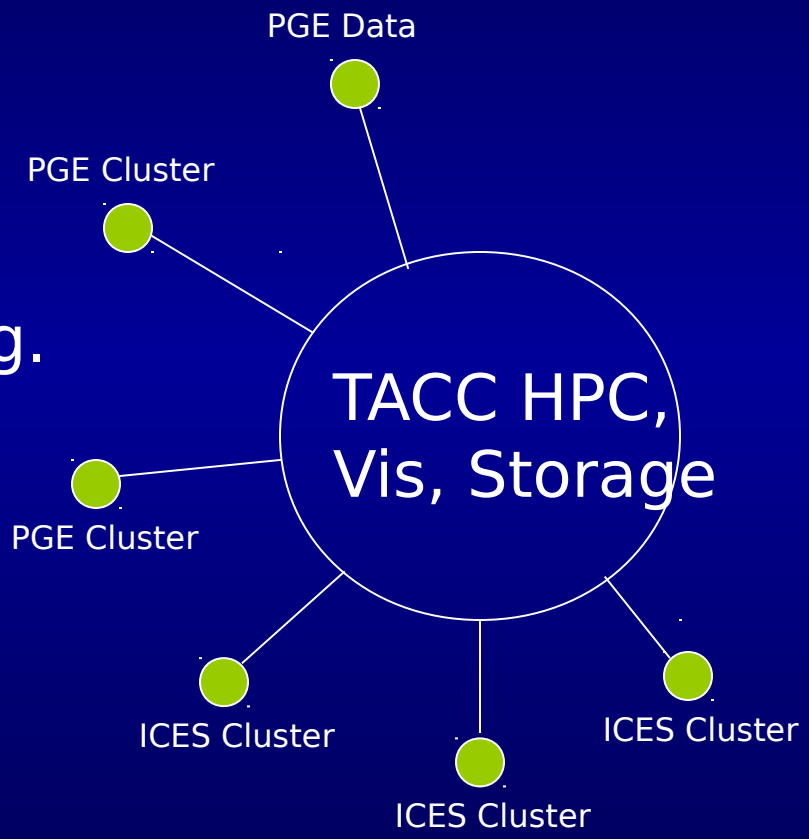
UT Grid: Logical View

- *Next* add other UT resources in one bldg. as spoke using *same tools and procedures*



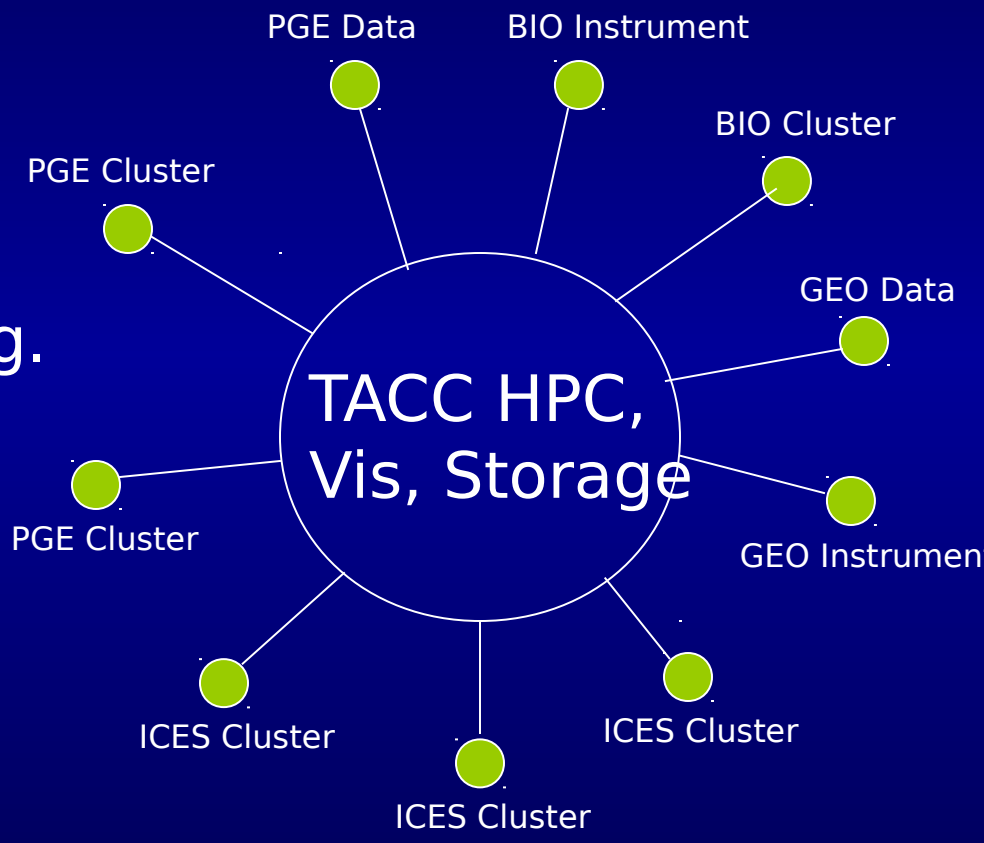
UT Grid: Logical View

- *Next* add other UT resources in one bldg. as spoke using *same tools and procedures*



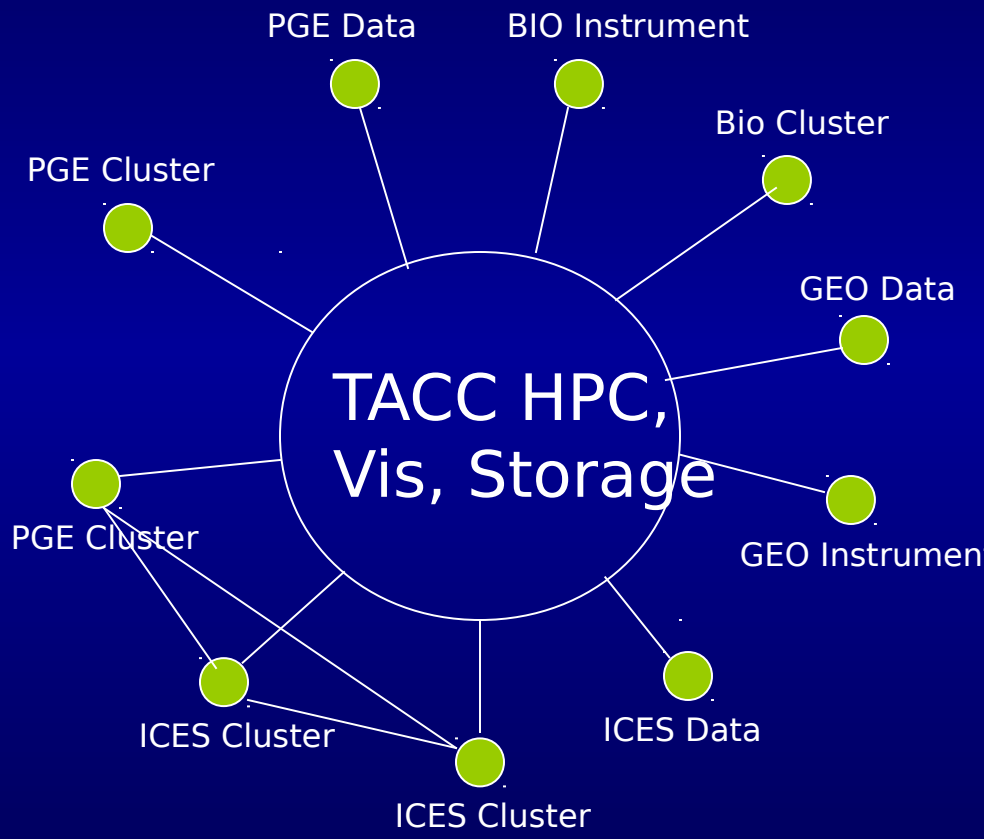
UT Grid: Logical View

- *Next* add other UT resources in one bldg. as spoke using *same tools and procedures*

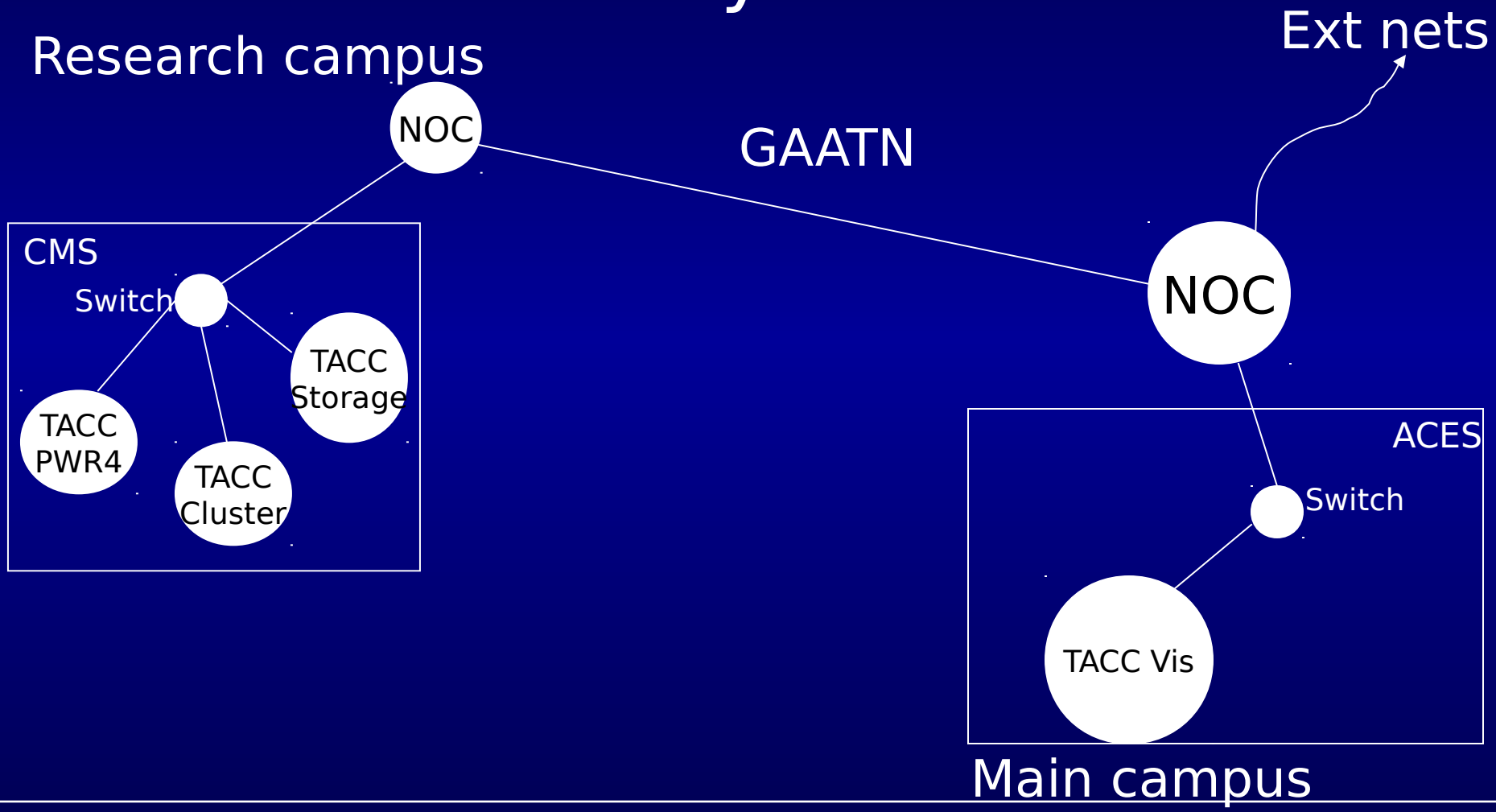


UT Grid: Logical View

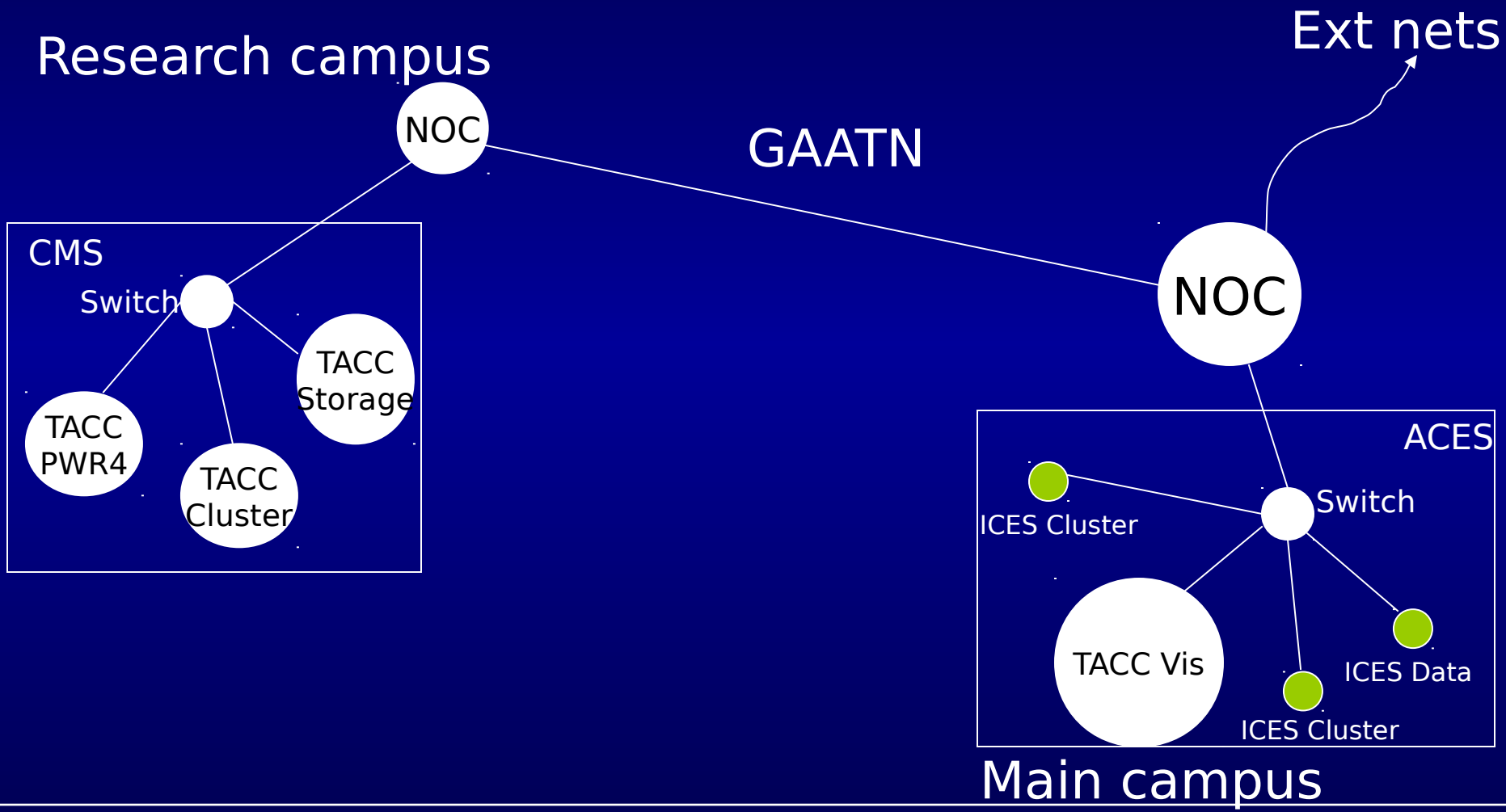
- *Finally* negotiate connections between spokes for *willing* participants to develop a P2P grid.



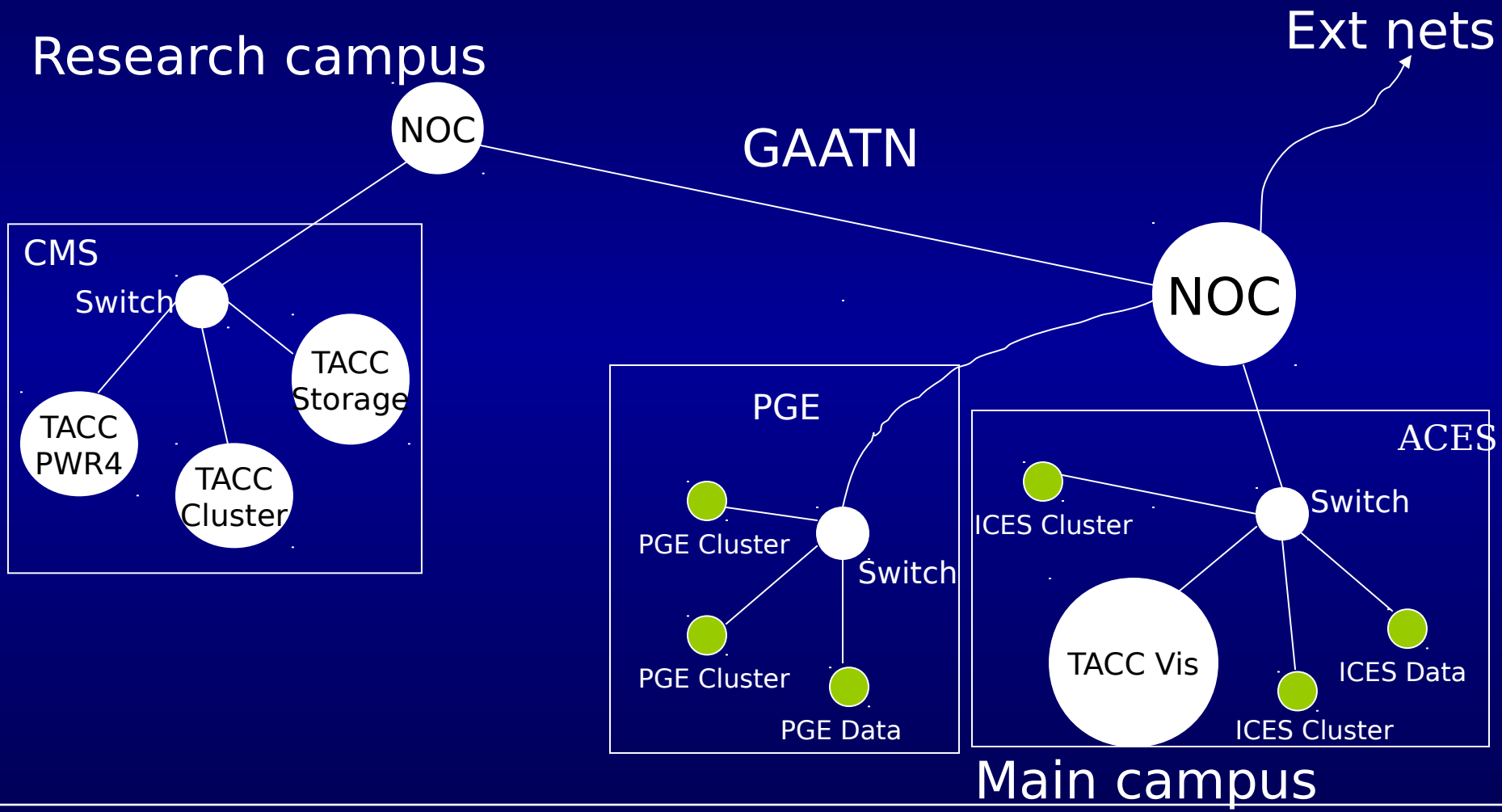
UT Grid: Physical View
TACC Systems



UT Grid: Physical View
Add ICES Resources



UT Grid: Physical View
Add Other Resources



UT Grid: Full Service!

- UT Grid will offer a complete set of services:
  - Compute services
  - Storage services
  - Data collections services
  - Visualization services
  - Instruments services
- But this will take 2 years—focusing on compute services now

UT Grid Interfaces

- Grid User Portal
  - Hosted, built on GridPort
  - Augment developers by providing info services
  - Enable productivity by simplifying production usage
- Grid User Node
  - Hosted, software includes GridShell plus client versions of all other UT Grid software
  - Downloadable version enables configuring local Linux box into UT Grid (eventually, Windows and Mac)

TACC Support of DoD CWO Community

PET Supports DoD HPC Research

- PET funds *external* experts for support of DoD research
  - high-level applications/resources support of DoD users
  - collaboration with DoD researchers
  - tech transfer into DoD research programs
  - development of technologies that enhance DoD research
- PET support must accelerate DoD research efforts, but must not serve as just programming labor
- DoD research benefiting from PET must be focused on use of HPCMPO resources

PET Supports DoD HPC Research

PET activities are based on functional areas (FAs)

DoD Computational Technology Area (CTA) FAs:	***Crosscutting Technology FAs:***
CWO – Climate/Weather/Ocean Modeling	CE – Computational Environments
EQM – Environmental Quality Modeling	ET – Enabling Technologies
CFD – Computational Fluid Dynamics	OKC – Online Knowledge Center
CSM – Computational Structural Mechanics	CDLT – Collaborative & Distance Learning Tech.
CCM – Computational Chemistry & Materials	EOT – Education, Outreach & Training
CEN – Computational Electronics & Nanotechnology	
CEA – Computational Electromagnetics & Acoustics	
SIP – Signal & Image Processing	
IMT – Integrated Modeling & Testing	
FMS – Forces Modeling & Simulation	

PET Supports DoD HPC Research

Each FA is supported by a:

- FA Point-of-Contact (FAPOC), from the PET team
- On-site staff in that FA, also from the PET team
- Component Point-of-Contact (CPOC), who oversees a group of three or four FAs, also from the PET team
- User Advisory Panel (UAP), comprising users of DoD HPCMPO systems

Contacts for PET CWO Support

- CWO On-sites
  - Tim Campbell (MSU@NAVO), tjcamp@navo.hpc.mil
  - John Romo (TACC/UT@Monterey), romo@tacc.utexas.edu
  - Roaming (based at TACC/UT)
    - Karl Schulz (TACC/UT), karl@tacc.utexas.edu
    - Tommy Minyard (TACC/UT, minyard@tacc.utexas.edu
- CWO FAPOC:
  - Jay Boisseau (TACC/UT), boisseau@tacc.utexas.edu
- CPOC for Component 1 (includes CWO, EQM, CE)
  - George Heburn (MSU@NAVO), heburn@navo.hpc.mil

Contacts for PET CWO Support

- Gov't Lead for Component 1: Eleanor Schroeder (NAVO)
- CWO CTA Leader: John Harding (CNMOC)
- MOS CWO User Advisory Panel
  - Rick Allard, NRL-SSC
  - Scott Applequist, AFWA
  - Cheryl Ann Blain, NRL-SSC
  - Mike Clancy, FNMOC
  - David Cooke, AFRL
  - Zeki Demirbilek, ERDC
  - Rich Hodur, NRL-Monterey
  - Julie McClean, NPS
  - Frank Ruggiero, AFRL

Strategic Objectives for CWO

- Improve the performance & scalability of strategic ocean and atmospheric models to enable higher resolution simulations and minimize time to produce forecasts.
- Improve the HPC user environment & tools to enable rapid porting and tuning of strategic CWO codes and of visualization/analysis of CWO data.
- Evaluate & develop new technologies & techniques in code coupling, data assimilation, ensemble calculations, and grids & coordinate systems for enhancing CWO simulation capabilities and accuracy.

DoD CWO User Requirements

1. Enhance strategic CWO codes: parallelizing, optimizing, & porting CWO codes for existing & new HPC systems
2. Easier-to-use programming tools & environments deployed consistently across DoD HPC systems
3. More scalable programming algorithms/solvers, tools, & systems
4. Code coupling between models

DoD CWO User Requirements

5. Enhanced capability for executing ensemble simulations: scheduling jobs, moving data, collecting results, etc.
6. Advanced CFD & mesh techniques for CWO
7. Enhanced visualization capabilities of model output for remote users
8. Acquisition & training for automatic adjoint compilers

PET Support Activities

1. Training
   - Requirements gathered by PET team
   - Taught by PET and 3rd parties (vendors, developers, etc.)
2. On-site staff
   - Based at major user sites
   - Provide support program-wide, not just at base user site
   - CWO has two half-time roaming on-sites
3. Projects
   - Formulated by PET team and external researchers
   - Reviewed by Government review panels

PET CWO Training

- Most CWO training requests are CE or ET
  - Requests have been used to schedule CE/ET classes at CWO sites
  - Requests have been used to formulate CE/ET classes with CWO slant
- Open to other topics, especially grid computing topics

PET CWO On-Site Activities

- Provide Technical Support
  - Provide technical support through on-site visits, phone, and email. High level support that goes beyond the “help desk”
- Porting Applications to HPC Platforms
  - Work with researchers to convert existing models from serial to parallel implementations
- Algorithm & Numerics Improvements
  - Work with researchers to investigate and implement algorithmic and numerical improvements to models

PET Projects

- Formulated once per year
- Can address specific issues more intensively than on-sites, training
- Please contact me if you have an idea for a project that would benefit your research

Summary

- PET supports HPC-related DoD user research on HPCMPO resources
- PET has developed a CWO support strategy identifying top user requirements
- PET activities to address user requirements:
  - On-sites
  - Projects
  - Training
- DoD CWO user requirements drive everything —need your input!

How Can We Help?

- Information sharing
  - Contact me about potential collaborations, or for contacts for specific grid computing projects
  - TACC employee here in Monterey, John Romo, is a conduit to rest of TACC
- Training
  - We teach classes in grid computing as well as HPC and SciVis (open to DoD users)
  - Can formulate custom grid computing classes for DoD CWO users under PET
- Projects
  - Can formulate PET project proposals to develop/deploy technologies that help you build a grid
  - Can formulate joint proposals to other funding sources for R&D

More About TACC:

Texas Advanced Computing Center

www.tacc.utexas.edu

info@tacc.utexas.edu

(512) 475-9411